A Lexicon Based Offline Arabic Handwritten Recognition Using Naive Bayesian Classifier with Gaussian Distribution

Ahlam Hasan Albashiti
Master of Informatics
Palestine Polytechnic University
Abu Roman St.,P.O.,Hebron,Palestine
ahlam@student.ppu.edu

Hashem Tamimi
Biotechnology Research Center
Palestine Polytechnic University
Abu Roman St.,P.O.198,Hebron,Palestine
htamimi@ppu.edu

May 23, 2012

Abstract

Offline handwritten recognition of Arabic script is difficult problem because of the overlapping between letters, each letter consists of one or more shapes according to its position in the word, in addition the existence of dots that change the meaning of the word. This leads to difficult analysis comparing with other languages. This paper introduces a lexicon based offline Arabic handwritten recognition using nave Bayesian classifier with Gaussian distribution.We report successful results for word, and part of words recognition.

Keywords Offline Arabic Handwritten Recognition, Naive Bayesian Classifier.

1 Introduction

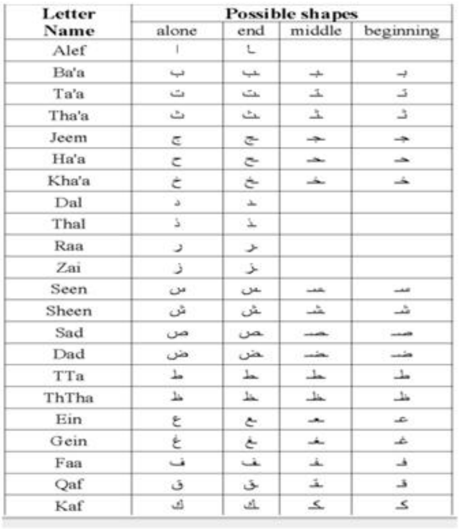
Handwriting recognition is used most often to describe the ability of a computer to translate human writing into text.[1]
Handwritten recognition can be either "online"[2, 3, 4] or"offline" [5, 6, 7, 8]according to the way in which the text is input to the computer. In online recognition the input is the sequence of points(X,Y)coordinates, the computer can reproduce the way of handwritten. On the other hand the input of the offline recognition is a 2D image. The online recognition is easier than the offline recognition because of the available information about the way of handwritten.[9, 10] This paper is restricted to the offline handwritten recognition for the Arabic script.

Outline In this paper we will give an overview about the Arabic language, then we will talk about the previous work of the Arabic handwritten recognition, after that we will talk about the methodology of the work,then we will show the results of the experiments, and finally the conclusions.

2 Overview of the Arabic Language

Arabic language consists of 28 characters, words are written from write to left in horizontal lines. Each character has more than one form depending on its position in the word [11]. See Figure 1.

Arabic handwritten is cursive, the characters in the word are joined together, this overlapping makes the recognition process difficult. Also some characters contains from one to three dots which change the meaning of the word [12]. See table 1.



Letter Name	Possible shapes			
	alone	end	middle	beginning
Alef	ا	ـا		
Ba'a	ب	ـب	ـبـ	بـ
Ta'a	ت	ـت	ـتـ	تـ
Tha'a	ث	ـث	ـثـ	ثـ
Jeem	ج	ـج	ـجـ	جـ
Ha'a	ح	ـح	ـحـ	حـ
Kha'a	خ	ـخ	ـخـ	خـ
Dal	د	ـد		
Thal	ذ	ـذ		
Raa	ر	ـر		
Zai	ز	ـز		
Seen	س	ـس	ـسـ	سـ
Sheen	ش	ـش	ـشـ	شـ
Sad	ص	ـص	ـصـ	صـ
Dad	ض	ـض	ـضـ	ضـ
TTa	ط	ـط	ـطـ	طـ
ThTha	ظ	ـظ	ـظـ	ظـ
Ein	ع	ـع	ـعـ	عـ
Gein	غ	ـغ	ـغـ	غـ
Faa	ف	ـف	ـفـ	فـ
Qaf	ق	ـق	ـقـ	قـ
Kaf	ك	ـك	ـكـ	كـ

Figure 1: Some Arabic characters and their forms according to their position in the word alone, end, middle and beginning of the word

Table 1: Arabic srripts

	A	B
1	Raa R	Zein Z
2	Raed	Zaied

3 Previous work

As we mentioned previously there are two approaches for recognition "offline" and "online", we will talk about them at the following.

3.1 Offline Recognition

The distinction between offline recognition methods is on the subject of segmentation of the word[13]. There are segmentation based methods and segmentation free methods, we will review both of them below.

First, segmentation based methods depend on segmenting the word into characters or into part of words in order to be recognized.

Bedda et al(Bedda et al,2006)[14]used NN classifier in order to classify 48 names of Algeria cities using 960 words(samples). The word is segmented into components, then each component is segmented into characters. For the classes each represents a shape where the same character can represents 1-4 classing according to their position in the word. For every class one definite a network neurons of type multilayered perceptions. The recognition rate for 960 word for the first writer is 98.33 ,and for the third writer is 90.

Ipson et al(Ipson,2009)used KMM classifier to recognize handwritten recognition, they segmented the word image into overlapping block then calculated the mean for each block, the result showed for 32492 words with 2 pixel overlapping block 76.042%. Second,Segmentation free based methods that depend on extracting global features from whole the image.[15, 16, 17].

(Somaya Almaadeed,2006) used NN classifier to recognize 70 different word, seven global features from whole the image are extracted,number of loops, position of ascender and descenders, lower dots, upper dots and the number of segments.

3.2 Online Recognition

In (El-Sana,2006) they took the sequence of (X,Y) coordinates for the word to extract three features from these point of sequences, for each point there are local angle, the angle between each point and the x-axis and the loop feature. According to the delayed stroke they developed the delayed stroke projection algorithm to detect the delayed stroke in the point of sequences and the incorporation of the delayed stroke in the word part body. After feature extraction, they used a discrete HMM to represent each letter shape. A network is built to represent each part of word, in this network each node is a letter and the path from node1 to node j represents part of word. A dictionary is used to recognize all the parts for a given word. For the evaluation the training data is built by four users ,each user wrote 800 word. For the testing data 2358 samples for ten users each of them 280 word The results are given according to the parts of words (segments) with respect to writer dependent and writer independent. And also according to the words for both writer dependent and writer independent samples. The recognition rate for part of words using 40 k dictionary 95.44% for writer dependent samples and 94.40%for writer independent. On the other hand 89.75% for words and writer dependent samples and 88.01% for writer independent samples.

In(Alimi,1997)[18] they build an online writer dependent system using neuro-fuzzy classifier, a genetic algorithm is used to select the best combination of letters that are recognized by the neuro-fuzzy classifier. In (Al-Emami,1990)[19] they developed an online system using decision tree classifier.

In(Zarka et.al,2009)[20] they developed an online Arabic handwritten recognition system based on new stroke segmentation algorithm, the proposed system gives an excellent recognition rate up to 97% and 92% for words and letter recognition.

4 Proposed Method

In this paper a recognition system is proposed where the main phases are included: preprocessing, feature extraction, classification, calculating probability, and word recognition . Once the sample image is acquired, preprocessing is required to enhance the image for better performance, after that, features are extracted for each part of the word (POW) using connected components features in the matlab ending up with feature matrix of 16 dimensions, then nave Bayesian classifier with Gaussian distribution is applied to decide to which class the unknown POW belongs, finally after calculating the probability of the combinations between POWs a lexicon is used to validate the given combination(word). Figure 2 shows the block diagram for the proposed work.

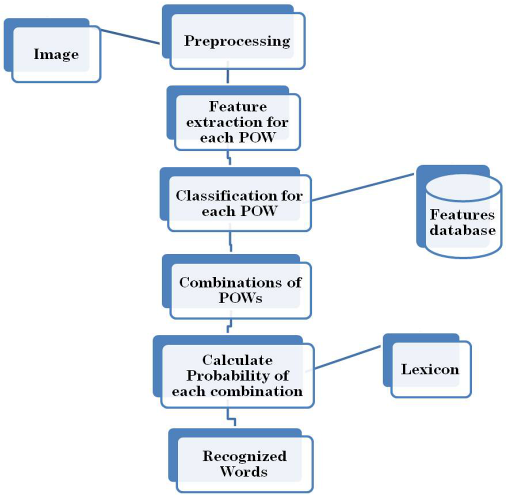


Image
Preprocessing
Feature extraction for each POW
Classification for each POW
Features database
Combinations of POWs
Calculate Probability of each combination
Lexicon
Recognized Words

Figure 2: Block Diagram of the work

1. **Preprocessing**
   Preprocessing of the image may include techniques such as normalization [21], thinning, baseline estimation[22], segmentation[13], skeletonizing [4]. Here the preprocessing includes image binarization, morphological operation for closing the spaces between pixels that are appeared from the user, after that word segmentation using connected component method in matlab. We segment each word into its POWs, according to the diacritical extraction step, we use a modified version algorithm proposed in[3], we assume each connected component with area less than 200 is a dot, In order to combine each diacritic to its connected component we choose the minimum distance between diaccritic and each connected component in the word.See table 2 the first figure is all the word, the second figure is the first POW of the word, and the last figure is the second POW of the word.

2. **Feature Extraction**
   The main objective of the feature extraction is to remove the redundant data and to produce a set of numerical features for the word image. These features are mapped to a classifier

Table 2: Connected Component Extraction in matlab and dots assigning. Dots are assigned to the second segment(minimum distance)

1) The input	2)POW1	3)POW2

to determine the corresponding class. In this work the feature vector of length 16 is computed using C.C (connected component region props in matlab). The features for each POW of the black-white word image are: the number of pixels of the region(area), center of the region(Xc,Yc), the length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region(minor axes length), the length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region(major axis length), the angle (in degrees ranging from -90 to 90 degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region (orientation), and finally the distance around the boundary of the region(perimeter), extrema - 8-by-2 matrix that specifies the extrema points in the region. Each row of the matrix contains the x- and y-coordinates of one of the points. The format of the vector is [top-left top-right right-top right-bottom bottom-right bottom-left left-bottom left-top].

3. **Classification**
   A multi-class classification system can be defined as follows: Given N-dimension space Ω, training data set Ω_{train} .$\Omega_{train} \in \Omega$ Each element x in Ω_{train} is associated with class label C where $C \in C_1....C_k$.For any tested feature vector x $f(x) \in$ class label C. [24]

   In this work the nave Bayesian classifier is used to map any feature vector to its corresponding class. In the next we will review this classifier.

   Nave Bayesian classifier is particularly suited when the dimensionality of the inputs is high. It is based on the Bayesian theorem [1].

$$P(c_j|x_1..x_n) = P(x_1..x_n|c_j)P(c_j) \qquad (1)$$

4. **Recognized Words**
   As we know nave Bayesian is a probability classifier, this means that we can rely on more than one result(recognized word) to be given to the user. the output of the naive Bayesian classifier is a matrix of size K for each POW contains probability for each class. In order to get the possible probability for the tested word, we make a combinations between all POWs then we find the probability for each combination according to the following. If the $word_i$ exists in the lexicon

$$word_{prob} = (P(POW_1)P(POW_2)...P(POW_h))*0.9 \qquad (2)$$

   otherwise

$$word_{prob} = (P(POW_1)P(POW_2)...P(POW_h))0.1 \qquad (3)$$

   where h is the number of POWs for a given word

5 Results

There is no reference data set for training and testing samples for the offline Arabic word recognition. Therefore we built our own data set for training and testing. According to the training data set we built 300 samples for 30 POW for two writers.For the testing we built 2 testing sets, first set consists of 29 words, the POW of these words are taken from the training data set.The second set test contains of POW that are not trained.

Table 3 shows the results of POWs where the features are connected components as mentioned in the above for different number of posteriors. Where table 4 shows the results of vertical-horizontal histogram for different posterities.
We noticed that connected component features gives

Table 3: Result Of POW for connected components features

Number of posteriors	Success rate
5	0.7679
6	0.7857
7	0.7946
8	0.8214
9	0.8482
10	0.8839

success rate higher than vertical-horizontal histogram.

Table 4: Result Of POW for vertical-horizontal histogram features

Number of posteriors	Success rate
5	0.3214
6	0.3304
7	0.3393
8	0.3839
9	0.4018
20	0.6786

Table 5 shows results of 29 tested words,where the POWs of each word are taken from the training data. Table 6 shows results of 29 tested words,where the POWs of each word not included in the training data. Another samples where added to be tested.

In our work the user can determine the best number of words that are smilax to the tested word, but this affect the time in seconds for classification. Table 7 gives the results for different number of posteriors(number of similar words).

Table 5: Result Of words from the POW that are exist in the training data

Number of posteriors	Success rate	Recognized word
1	6786	19
5	0.9286	26
8	0.9287	26

Table 6: Result Of words from the POW that are not exist in the training data

Number of posteriors	Success rate	Recognized word
1	.4286	12
3	0.7679	19
5	0.7857	22
8	.7857	22

6 Conclusions

In this paper we build a lexicon based offline Arabic handwritten recognition using nave Bayesian classifier with Gaussian distribution, our contribution is not only to guess the word, also to give the user the most similar words to the tested word depending on the data from the lexicon. the results reported the success of the method.

References

[1] M. M. HAJI, "Farsi handwritten word recognition using continuous hidden markov models and structural features," Master's thesis, SHIRAZ UNIVERSITY SHIRAZ, IRAN, 2005.

[2] C. E. VIARD-GAUDIN, "Using segmentation constraints in an implicit segmentation scheme for on-line word recognition," in *International Workshop on Frontiers in Handwriting Recognition*, 2006.

Table 7: Results for 29 tested word with respect to time for differen posteriors

Number of posteriors	Time
1	16.0444
3	16.0787
5	16.0745
8	16.1050
10	16.24

[3] A. R. K. M. V.-G. C. P. E, "Online handwriting recognition using support vector machine," *IEEE Region 10 Conference*, 2004.

[4] A. A.M., "An evolutionary neuro-fuzzy approach to recognize on-line arabic handwriting," in *Document Analysis and Recognition, Proceedings of the Fourth International Conference on*, pp. 382 – 386, 1997.

[5] S. Alma'adeed, "Recognition of off-line handwritten arabic words using neural network," in *Geometric Modeling and Imaging–New Trends*, pp. 141–144, 2006.

[6] T. Klassen, "Towards neural network recognition of handwritten arabic letters," Master's thesis, Dalhousie University, 2001.

[7] M. Pechwitz and V. Maergner, "Hmm based approach for handwritten arabic word recognition using the ifn/enit - database," *Proceedings of the Seventh International Conference on Document Analysis and Recognition IEEE.*, 2003.

[8] Khorsheed, "Recognising handwritten arabic manuscripts using a single hidden markov model pattern recognition letters," in *Geometric Modeling and Imaging–New Trends*, 2003.

[9] L. M. Lorigo and V. Govindaraju, "Offline arabic handwriting recognition:a survey," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2006.

[10] B. Alsallakh and H. Safadi, "Arapen: an arabic online handwriting recognition system," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2006.

[11] A. Amin, "offline arabic character recognition," in *pattern recognition*, 1998.

[12] A. T. Al-Taani and S. Al-Haj, "Recognition of on-line arabic handwritten characters using structural features," *JOURNAL OF PATTERN RECOGNITION RESEARCH*, 2010.

[13] P. Burrow, "Arabic handwriting recognition," Master's thesis, University of Edinburgh, 2004.

[14] M. SEPTI and M. BEDDA, "Contribution to the recognition of hand arabic word based on neural network," *IEEE International Conference on Signal and Image Processing Applications*, 2006.

[15] V. Govindaraju and R. K. Krishnamurthy, "Holistic handwritten word recognition using temporal features derived from off-line images," *Pattern Recognition*, 1996.

[16] D. Gorsky, "Experiments with handwriting recognition using holographic representation of line images," *Pattern Recognition*, 1994.

[17] C. Parisse, "Global word shape processing in off-line recognition of handwriting," *IEEE Trans. Pattern Anal. Mach. Intell*, 1994.

[18] A. M. Alimi, "An evolutionary neuro-fuzzy approach to recognize on-line arabic handwriting," in *International Conference Document Analysis and Recognition*, 199.

[19] S. Al-Emami and M. Usher, "On-line recognition of handwritten arabic characters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1990.

[20] K. Daifallah, N. Zarka, and H. Jamous, "Recognition-based segmentation algorithm for on-line arabic handwriting," in *International Conference on Document Analysis and Recognition*, 2009.

[21] A. Benouareth, A. Ennaji, and M. Sellami, "Hmms with explicit state duration applied to handwritten arabic word recognition," *ICPR*, pp. 897–90, 2006.

[22] A. J and R. J, "Knowledgebased baseline detection and optimal thresholding for words segmentation in efficient pre-processing of handwritten arabic text," in *Proc. 5th Int. Conf. Information Technology:New Generation*, pp. 1158–1159, 2008.

[23] Bozinovic and Srihari.S, "offline cursive script word recognition," *IEEE Trans on PAMI*, 1989.

[24] J. H. AIKhateebl, F. Khelifil, J. Jiani, and S. S. Ipsonl, "A new approach for off-line handwritten arabic word recognition using knn classifier," *IEEE International Conference on Signal and Image Processing Applications*, 2009.

[25] S. Almaadeed, "Recognition of off-line handwritten arabic words using neural network," *IEEE International Conference on Signal and Image Processing Applications*, 2006.

[26] F. Biadsy, J. El-Sana, and N. Habash, "Online arabic handwriting recognition using hidden markov models," *HAL-INRIA*, 2006.

[27] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996.

[28] S. S. El-Dabi, R. Ramsis, and A. Kuwait, "Arabic character recognition system: a statistical approach for recognizing cursive typewritten text," *Pattern Recognition*, 1990.

[29] B. R and S. S, "Off-line cursive script word recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989.

[30] V. Mrgner, H. E. Abed, and M. Pechwitz, "Offline handwritten arabicword recognition using hmm a character based approach without explicit segmentation," *Institut for Communications Technology*, pp. 897–90, 2006.